DOCUMENT RESUME

ED 137 343 TM 006 144

AUTHOR Macready, George B.; Dayton, C. Mitchell
TITLE Statistical Comparisons Among Hierarchies Based on Latent Structure Models. Research Monograph 77-1.
INSTITUTION Maryland Univ., College Park. Dept. of Measurement and Statistics.
PUB DATE Apr 77
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Goodness of Fit; *Hypothesis Testing; *Mathematical Models; Probability; Role Conflict; Standard Error of Measurement; *Statistical Analysis; Tests of Significance; True Scores
IDENTIFIERS Domain Referenced Tests; *Latent Structure Analysis

ABSTRACT

A probabilistic hypothesis testing procedure to assess the fit of hypothesized hierarchical structures for test item data is discussed. Statistical procedures are presented which are useful for evaluating the fit of data of a certain class of probabilistic models. These models apply to sets of dichotomous (0,1) responses for which there are posited to exist a priori dependence structures. Examples of relevant types of data are success/failure patterns from Piagetian tasks, learning hierarchies, and domain referenced tests, as well as agree/disagree responses from attitude tests. (Author/RC)

MEASUREMENT and STATISTICS

COLLEGE OF EDUCATION
UNIVERSITY OF MARYLAND

STATISTICAL COMPARISONS AMONG HIERARCHIES BASED ON LATENT STRUCTURE MODELS

by

George B. Macready

&

C. Mitchell Dayton

April 1977

Department of Measurement and Statistics
College of Education
University of Maryland

I. THEORY

Introduction

The purpose of this paper is to present statistical procedures which are useful for evaluating the fit to data of a certain class of probabilistic models. These models apply to sets of dichotomous (0,1) responses for which there are posited to exist a priori dependency structures. Examples of relevant types of data are success/failure patterns from Piagetian tasks, learning hierarchies, and domain-referenced tests, as well as agree/disagree responses from attitude instruments.

Summary of the Model

Using the notation from Dayton and Macready (1976), where the model is developed in more detail, we assume K distinct tasks each of which can be scored 0,1 for a sample of n individuals. Corresponding to an a priori dependency structure, a hypothetical set of 0,1 response patterns (true score patterns) exists which would typify an "ideal" group of respondents (i.e., a group matching the latent structure). We let

$$(1) \qquad P(\underline{u}_s) = \sum_{j=1}^{q} P(\underline{u}_s|\underline{v}_j)\theta_j$$

be the probability of an observed response vector, $\underline{u}_s$, where there are q hypothetical true score patterns, $\underline{v}_j$, $j = 1,\ldots,q$, with relative frequencies of occurrence, θ_j ($\sum_{j=1}^{q} \theta = 1$). The conditional probabilities, $P(\underline{u}_s\ \underline{v}_j)$ are "recruitment" probabilities which connect the observed response patterns to the true score patterns. The general class of recruitment probabilities which are of interest take on the form:

$$(2) \qquad P(\underline{u}_s|\underline{v}_j) = \prod_{i=1}^{K} \alpha_i^{a_{ij}}(1-\alpha_i)^{b_{ij}}\beta^{c_{ij}}(1-\beta_i)^{d_{ij}} .$$

The parameters α_i and β_i are "error" probabilities which are interpretable, respectively, as "intrusion" (e.g., guessing) and "omission" (e.g., forgetting) error rates, while the coefficients, a_{ij} through d_{ij} are 0 or 1 and are chosen to correspond to the particular $\underline{u}_s$ and $\underline{v}_j$ vectors involved.

For example, with K = 3 and an a priori Guttman scale, the 4 true score patterns would be $\underline{v}_1 = (0\ 0\ 0)'$, $\underline{v}_2 = (1\ 0\ 0)'$, $\underline{v}_3 = (1\ 1\ 0)'$, and $\underline{v}_4 = (1\ 1\ 1)'$, and the different possible observed vectors, $\underline{u}_s$, are the $2^3 = 8$ ordered sets of 0's and 1's: $(0\ 0\ 0)'$; $(1\ 0\ 0)'$; $(0\ 1\ 0)'$; $(1\ 1\ 0)'$; $(0\ 0\ 1)'$; $(1\ 0\ 1)'$; $(0\ 1\ 1)'$; $(1\ 1\ 1)'$. Each of the 8 possible observed vectors may arise from any one of the 4 true score patterns by suitable choices for the coefficients a_{ij} through d_{ij}. For simplicity, let $\bar{\alpha}_i = 1 - \alpha_i$ and $\bar{\beta}_i = 1 - \beta_i$; then, the recruitment probabilities, $P(\underline{u}_s|\underline{v}_j)$ are:

Observed Pattern	True Score Patterns 0 0 0	1 0 0	1 1 0	1 1 1
0 0 0	$\bar{\alpha}_1\bar{\alpha}_2\bar{\alpha}_3$	$\beta_1\bar{\alpha}_2\bar{\alpha}_3$	$\beta_1\beta_2\bar{\alpha}_3$	$\beta_1\beta_2\beta_3$
1 0 0	$\alpha_1\bar{\alpha}_2\bar{\alpha}_3$	$\bar{\beta}_1\bar{\alpha}_2\bar{\alpha}_3$	$\bar{\beta}_1\beta_2\bar{\alpha}_3$	$\bar{\beta}_1\beta_2\beta_3$
0 1 0	$\bar{\alpha}_1\alpha_2\bar{\alpha}_3$	$\beta_1\alpha_2\bar{\alpha}_3$	$\beta_1\bar{\beta}_2\bar{\alpha}_3$	$\beta_1\bar{\beta}_2\beta_3$
1 1 0	$\alpha_1\alpha_2\bar{\alpha}_3$	$\bar{\beta}_1\alpha_2\bar{\alpha}_3$	$\bar{\beta}_1\bar{\beta}_2\bar{\alpha}_3$	$\bar{\beta}_1\bar{\beta}_2\beta_3$
0 0 1	$\bar{\alpha}_1\bar{\alpha}_2\alpha_3$	$\beta_1\bar{\alpha}_2\alpha_3$	$\beta_1\beta_2\alpha_3$	$\beta_1\beta_2\bar{\beta}_3$
1 0 1	$\alpha_1\bar{\alpha}_2\alpha_3$	$\bar{\beta}_1\bar{\alpha}_2\alpha_3$	$\bar{\beta}_1\beta_2\alpha_3$	$\bar{\beta}_1\beta_2\bar{\beta}_3$
0 1 1	$\bar{\alpha}_1\alpha_2\alpha_3$	$\beta_1\alpha_2\alpha_3$	$\beta_1\bar{\beta}_2\alpha_3$	$\beta_1\bar{\beta}_2\bar{\beta}_3$
1 1 1	$\alpha_1\alpha_2\alpha_3$	$\bar{\beta}_1\alpha_2\alpha_3$	$\bar{\beta}_1\bar{\beta}_2\alpha_3$	$\bar{\beta}_1\bar{\beta}_2\bar{\beta}_3$

The total probability for a given observed pattern is the weighted sum of the appropriate recruitment probabilities using the weights, Θ_j. E.g.,

$$P(\underline{u}_s = (0\ 1\ 0)') = \Theta_1\bar{\alpha}_1\alpha_2\bar{\alpha}_3 + \Theta_2\beta_1\alpha_2\bar{\alpha}_3 + \Theta_3\beta_1\bar{\beta}_2\bar{\alpha}_3 + \Theta_4\beta_1\bar{\beta}_2\beta_3$$

Estimation Procedures

For $n = \sum_{s=1}^{2^K} n_s$ respondents, the likelihood for the sample is

$$L = \prod_{s=1}^{2^K} P(\underline{u}_s)^{n_s} \tag{3}$$

where $\underline{u}_s$ is an observed 0,1 vector given by n_s respondents. With q a priori true score patterns, there are $2K + q - 1$ independent parameters to estimate and problems arise from 2 sources: (I) the parameters may be non-identifiable;

(II) the set of partial derivatives of L with respect to the θ_j , α_i , and β_i (the normal equations) are, in general, non-linear in the parameters.

(I) Non-Identifiable Models - consistent estimates will not be available in this circumstance and the model must be restricted suitably to permit estimation. For example, with true score patterns of $\underline{v}_1 = (0\ 0\ \cdots\ 0)'$ and $\underline{v}_2 = (1\ 1\ \cdots\ 1)'$ only, the model is identifiable so long as $K \geq 3$. However, with true score patterns typifying a linear hierarchy, the model is not identifiable; the restrictions $\alpha_i = \alpha$ and $\beta_i = \beta$ do result in an identifiable model so long as $K \geq 4$.

(II) Non-Linear Normal Equations - since the data may be represented as frequencies of occurrence, n_s , for the 2^K possible 0,1 outcome patterns, an iterative maximum likelihood (ML) estimation scheme (Fisher's method of scoring) can be employed (Rao, 1965). Computer programs written in FORTRAN IV have been developed around this iterative ML algorithm for the following cases[1]:

MODEL3 - true score patterns are $(0\ 0\ \cdots\ 0)$ and $(1\ 1\ \cdots\ 1)$ only: $\alpha_i = \alpha$, $\beta_i = \beta$ is assumed (this model is, simply, a mixture of two binomials).

MODEL3G - true score patterns are $(0\ 0\ \cdots\ 0)$ and $(1\ 1\ \cdots\ 1)$ only; α_i and β_i are estimated per task (item).

MODEL5 - true score patterns may be any linear or branching hierarchy; $\alpha_i = \alpha$ and $\beta_i = \beta$ are assumed (optionally, $\alpha_i = \beta_i = \alpha$ can be imposed as a further restriction).

Assessing Fit of the Model

Standard (Pearson) chi-square goodness-of-fit tests can be utilized by computing "expected" frequencies for each of the 2^K possible response patterns. Let

$$\hat{P}(\underline{u}_s) = \sum_{j=1}^{q} \hat{P}(\underline{u}_s|\underline{v}_j)\cdot\hat{\theta}_j \qquad (4)$$

be found by substituting ML estimates in (1) and (2). Then, the expected

[1]Single copies of program listings and a user's manual are available by writing the authors at Department of Measurement & Statistics, College of Education, University of Maryland, College Park, Md. 20742.

frequencies are given by

$$\hat{n}_s = n \cdot \hat{P}(\underline{u}_s) \tag{5}$$

and the Pearson chi-square statistic is

$$c_p = \sum_{s=1}^{2^K} [(n_s - \hat{n}_s)^2/\hat{n}_s] \tag{6}$$

which can be evaluated (for "large" n) as chi-square with degrees of freedom equal to $2^K - q' - 1$, where q' is the number of independent parameters estimated under the model (e.g., $q' = 3$ in MODEL3; $q' = 2K + 1$ in MODEL3G; $q' = q + 1$ in MODEL5).

As an alternative to the Pearson chi-square statistic, the fitted model can be compared to the best-fitting multinomial density by a likelihood ratio test. The estimators for $P(\underline{u}_s)$ under the multinomial model are n_s/n and the likelihood ratio is

$$\lambda = \prod_{s=1}^{2^K} [\hat{P}(\underline{u}_s)/(n_s/n)]^{n_i} \tag{7}$$

where $\hat{P}(\underline{u}_s)$ is as defined in (4). For "large" n, $c_L = -2\log_e\lambda$ is a chi-square statistic with $2K - q' - 1$ degrees of freedom (the Pearson and likelihood ratio statistics are asymptotically equivalent).

Comparisons among Models

Two different forms of the probabilistic model in (1) and (2) can be compared on the same set of data if one model can be derived from the other by imposing linear restrictions on the parameters. For example, MODEL3 can be derived from MODEL3G by setting $\alpha_i = \alpha$ and $\beta_i = \beta$; thus, the relative fits of these two models can be compared. Similarly, there is great flexibility in comparing different hierarchic structures under MODEL5. Models related in the above manner are described as exhibiting "subset inclusion" among the parameters. By an extention of the likelihood ratio test for fit in (7), we can compare models obeying subset inclusion. Assume that the more complex model is based on

fitting r_1 parameters, while the less complex model involves $r_2 < r_1$ parameters; that is, $r_1 - r_2$ restrictions have been imposed when deriving the second model from the first model. Let λ_1 and λ_2 be likelihood ratios derived as in (7) for the respective models. Than, $c_{12} = -2\log_e(\lambda_2/\lambda_1) = c_2 - c_1$ is a chi-square statistic with $r_1 - r_2$ degrees of freedom and this statistic provides a basis for deciding whether or not the more restricted form of the model is a poorer fit to the data than the more complex model.

Cross-Validation of Models

The same form of the probabilistic model (e.g., linear hierarchy) may be posited for samples which differ systematically (e.g., males and females). Although some general procedures can be used to compare the consistency of observed frequencies in different samples (chi-square or Kolmogorov-Smirnov statistics), the comparability of parameter estimates can be assessed by a double cross-validation technique. That is, parameter estimates for the relevant parameters (Θ_j , α_i , etc.) are derived from each sample separately and then fitted to frequencies from the other sample. With appropriate modifications to degrees of freedom ($2^K - 1$ rather than $2^K - q' - 1$), the cross-validation chi-squares for goodness-of-fit provide evidence for the consistency of parameter estimates across samples.

Significance Tests for Parameters

Inter-sample and intra-sample significance tests are available for individual parameters (assuming large samples) since the interative ML estimation procedure yields asymptotic sampling variance-covariance estimates. Appropriate intra-sample hypotheses are $\Theta_j = D_j$, $\alpha_i = A_i$, or $\beta_i = B_i$, where D_j, A_i, and B_i are ordinarily 0, and the test statistics are

(8) $$z = (\hat{\Theta}_j - D_j)/S_{\Theta_j} \quad , \quad z = (\hat{\alpha}_i - A_i)/S_{\alpha_i} \quad , \text{ etc.}$$

If several such tests are conducted for the same set of data and simultaneous control of the Type I error rate is desired, the Bonferroni (Fisher-Dunn) approach

is generally appropriate and involves merely the setting of the significance level per test at 1/m of the total desired Type I error rate (where m is the number of statistical tests being conducted). In addition, it is possible to test hypotheses based on alleged relationships among subsets of the parameters. A common example involves MODEL5 where the equality of intrusion and omission error rates would imply the hypothesis $\alpha = \beta$. Since "large" sample estimates of the relevant variances and covariances are available, the test can be set up as

$$z = (\hat{\alpha} - \hat{\beta}) / \sqrt{S_{\alpha}^2 + S_{\beta}^2 - 2S_{\alpha\beta}} \quad . \tag{9}$$

Similarly, inter-sample tests for hypotheses such as $\alpha_{i1} = \alpha_{i2}$ are of the form

$$z = (\hat{\alpha}_{i1} - \hat{\alpha}_{i2}) / \sqrt{S_{\alpha_{i1}}^2 + S_{\alpha_{i2}}^2} \quad , \tag{10}$$

where the sample is referenced by the second subscript (e.g., $\hat{\alpha}_{i1}$ is the estimate for the "guessing" error rate on task i in sample 1).

II. APPLICATIONS OF MODELS

In this section of the paper two sets of data are used as the basis of separate analyses in order to provide examples of a variety of analytic procedures that can be applied within the context of the models.

Role Conflict Example

The first example is based on the data from a study by Stouffer and Toby (1951) dealing with individuals role conflict in determining "the proper thing to do" in a morally conflicting situation involving conflicts between obligations to a friend and more general social obligations.

Their data are based on two forms of a four item questionnaire both of which were completed by 216 randomly assigned undergraduate students.

For form I (Ego faces dilemma) of the questionnaire, the respondent was faced with the following role conflicts:

1. You are riding in a car driven by a close friend, and he hits a a pedestrian. You know he was going at least 35 miles an hour in a 20-mile-an-hour speed zone. There are no other witnesses. His lawyer says that if you testify under oath that the speed was only 20 miles an hour, it may save him from serious consequences. What do you think you'd probably do in view of the obligations of a sworn witness and the obligation to your friend?

   Check one:

   _____ Testify that he was going 20 miles an hour
   _____ Not testify that he was going 20 miles an hour.

2. You are a New York drama critic. A close friend of yours has sunk all his savings in a new Broadway play. You really think the play is no good. Would you go easy on his play in your review in view of your obligations to your readers and your obligation to your friend?

   Check one:

   _____ Yes
   _____ No

3. You are a doctor for an insurance company. You examine a close friend who needs more insurance. You find that he is in pretty good shape, but your are doubtful on one or two minor points which are difficult to diagnose. Would you shade the doubts in his favor in view of your obligations to the insurance company and your obligation to your friend?

   Check one:

   _____ Yes
   _____ No

4. You have just come from a secret meeting of the board of directors of a company. You have a close friend who will be ruined unless he can get out of the market before the board's decision becomes known. You happen to be having dinner at that friend's home this same evening. Would you tip him off in view of your obligations to the company and your obligation to your friend?

Check one:

_____ Yes
_____ No

While for form II (Friend faces dilemma) of the questionnaire the stories were rewritten so that a friend of the respondent was faced with the same dilemmas.

On the basis of a Guttman scalogram analysis Stouffer & Toby (1951) posite that there may be a linear scale underlying their instrument. They state:

> This fusion of variables in our situation does seem to generate a unidimensional scale, the dimension involved being the degree of strength of a latent tendency to be loyal to a friend even at the cost of other principles. The rank groupings would represent ordered degrees of probability of taking the friend's side in a role conflict p. 400.

The Guttman scalogram analysis, for both questionnaires resulted in the following order of items: 4, 3, 2, 1 where all preceding items are considered to be conditional prerequisites for responding positively to an item. This ordering resulted in reproducibility coefficients of .92 and .91, respectively, for forms I and II. These values are both larger than the minimally sufficient value of .90 suggested by Guttman as necessary for a linear scale (Torgerson, 1958). However, as Stouffer & Toby point out, there are two response patterns (1 1 0 1 and 1 0 1 0 for items 4 through 1, respectively, where a "1" indicates a yes response to an item and a "0" indicates a no response to an item) with relatively high frequencies of occurrence which are not compatible with the linear scale (see Tables I & II).

If these two response patterns are added as "true score" response patterns to those true score patterns for a linear scale (see footnotes to Tables I and II) a resulting "branching hierarchy" (as described in Macready, 1975) is obtained in which the same conditional relations are present as for the posited linear scale

Table I

Response Frequencies and Tests of Model Fit for Role Conflict Data-Form I

Response Patterns	Observed Frequencies	Expected Frequencies	
Items: 4 3 2 1		Linear Model	Branching Hierarchy
0 0 0 0[a,b]	42	41.057	41.400
1 0 0 0[a,b]	23	24.102	23.443
0 1 0 0	6	8.248	5.903
0 0 1 0	6	5.640	6.138
0 0 0 1	1	1.899	1.742
1 1 0 0[a,b]	24	22.169	24.002
1 0 1 0[b]	25	14.117	25.214
1 0 0 1	4	2.567	2.565
0 1 1 0	7	13.608	7.335
0 1 0 1	2	2.058	1.973
0 0 1 1	1	1.974	.899
1 1 1 0[a,b]	38	41.909	37.573
1 1 0 1[b]	9	6.249	9.947
1 0 1 1	6	5.990	4.417
0 1 1 1	2	5.974	3.813
1 1 1 1[a,b]	20	18.441	19.637

Reproducibility Coefficient	.92		
Chi Square Tests			
Goodness of Fit	18.5657		2.7684
Difference in Fit		15.7973	
Degrees of Freedom	9	2	7
P-Value	.029	.000	.906

[a]True score response patterns for the linear scale model.
[b]True score response patterns for the branching hierarchy model.

Table II

Response Frequencies and Tests of Model Fit for Role Conflict Data-Form II

Response Patterns	Observed Frequencies	Expected Frequencies	
Items: 4 3 2 1		Linear Model	Branching Hierarchy
0 0 0 0[a,b]	37	29.783	37.028
1 0 0 0[a,b]	31	31.679	30.688
0 1 0 0	5	11.321	6.810
0 0 1 0	6	6.629	4.520
0 0 0 1	2	5.948	2.270
1 1 0 0[a,b]	29	32.895	27.570
1 0 1 0[b]	15	10.209	16.281
1 0 0 1	4	6.918	5.111
0 1 1 0	6	6.251	5.256
0 1 0 1	3	2.960	4.248
0 0 1 1	3	2.048	.967
1 1 1 0[a,b]	25	25.975	25.673
1 1 0 1[b]	23	10.064	20.701
1 0 1 1	4	5.653	4.530
0 1 1 1	3	4.884	4.094
1 1 1 1[a,b]	20	22.783	20.173
Reproducibility Coefficient		.91	
Chi Square Tests			
Goodness of Fit		27.9201	5.3686
Difference in Fit		22.5515	
Degrees of Freedom		9 2	7
P-Value		.001 .000	.615

[a]True score response patterns for the linear scale model.
[b]True score response patterns for the branching hierarchy model.

except that:

(a) item 3 is not a conditional prerequisite for item 2, and

(b) item 2 is not a conditional prerequisite for item 1.

If it is assumed that $\alpha_i = \alpha$ and $\beta_i = \beta$ for $i=(1,2,3,4)$ then both the linear model and the branching hierarchy described above are special cases for Model 5 described in part I of this paper. The resulting expected frequencies for the response patterns (based on Maximum Likelihood parameter estimates) under each of the above models are presented in Tables I & II, for questionnaire forms I & II respectively. Note that the accuracy of the estimated frequencies for the branching hierarchy in most cases provides closer approximations to the observed frequencies than those obtained under the linear model. As might be expected, chi-square tests of fit for each test form with level of significance set at .05 resulted in "acceptable fit" only for the branching hierarchy. In addition, the branching hierarchy model was found to provide significantly better fit than was provided by the linear model (see Tables I & II).

Equality between corresponding parameters under the branching hierarchy model for the two-questionnaire forms were simultaneously tested via a double cross-validation procedure. This analysis, the results of which are presented in Table III, led to the rejection of the hypothesis of equality when a .05 level of significance was used. This is supportive evidence for separate post hoc comparisons testing equality of values for each parameter found under the two forms.

Table III

Double Cross-validation for the Hierarchic Model across Forms of Questionnaire

Questionnaire Form		Chi Square	df	P-Value
Parameter estimates	Fitted data			
A	B	30.600	15	.010
B	A	26.305	15	.035

Table IV

Maximum Likelihood Parameter Estimates and their Standard Errors for Role Conflict Data

True Score Patterns	Form I				Form II				Questionnaire Form Parameter Comparisons for Hierarchic Model	
	Linear Model		Hierarchic Model		Linear Model		Hierarchic Model			
Items: 4 3 2 1	Parameter est.	Std. Error	Parameter est.	Std. error	Parameter est.	Std. Error	Parameter est.	Std. error	z scores	2-tailed P-Values
0 0 0 0	.1765	.041	.1961	.039	.2281	.062	.1679	.042	.49	.62
1 0 0 0	.1080	.036	.0871	.031	.2019	.052	.1376	.037	-1.03	.30
1 1 0 0	.0703	.040	.1072	.034	.2274	.052	.1331	.037	- .52	.60
1 1 1 0	.3995	.058	.2678	.046	.1611	.049	.1724	.040	1.57	.12
1 1 1 1	.2457	.051	.1722	.043	.1815	.055	.1809	.047	- .14	.89
1 0 1 0	---	---	.1236	.034	---	---	.0742	.029	1.12	.26
1 1 0 1	---	---	.0460	.024	---	---	.1339	.033	-2.15	.03
Response Errors										
α	.0311	.031	.0327	.029	.1628	.039	.0380	.037	- .11	.91
β	.2446	.031	.1625	.034	.1714	.045	.1686	.036	- .13	.90

The maximum likelihood parameter estimates and their corresponding estimated standard errors obtained under each of the described models for forms I & II are presented in Table IV. Note that under the hierarchic model, there are "moderate" (relative to the standard errors) differences found between the form estimates of corresponding proportions for some of the true score response patterns (namely 1 0 0 0, 1 1 1 0, 1 0 1 0 and 1 1 0 1). However, a significant difference between the estimated proportions occured only in the case of response pattern 1 1 0 1. On the other hand, corresponding estimates of each of the error parameters show extremely small differences. These combined findings provide support for the contention that differences that do exist for the two procedures of testing, do not affect "error rates" but do affect the proportion of individuals found within each of the true score response patterns. The specific nature of this effect is, however, at best vague.

Based on the differences under the hierarchic model in the estimated true score proportions of individuals who "should" respond positively to items 1 through 4, for forms I & II (these differences are respectively -.09, +.13, -.03 and -.03) the following conjecture seems appropriate: individuals under questionnaire form I tend to produce more simultaneous positive responses on items 1 and 3, which are in one branch of the posited hierarchy, while individuals under questionnaire form II tend to produce more positive responses to item 2, which is at the end of other branch of the hierarchy.

Domain Referenced Testing Example

The second example is based on the data from a study by Macready and Dayton (1976) in which the relations among items from a single domain in a Domain Referenced Test, (DRT), were investigated. This domain contains items involving integer multiplication in which: (a) the multiplier has 2 digits; (b) the multiplicand has either 3 or 4 digits and (c) there is at least one "carry" operation for each digit in the multiplier. The specific data considered is based on dicotomus item

scores (i.e., the scores 1 & 0 indicate respectively passing and failing the item in question) obtained on 4 randomly selected items from the specified domain for 284 fourth grade students.

Macready and Merwin (1973) as well as Harris (1974) have suggested that the construction and revision of item domains be based on the homogeneity of item content and the internal consistency of examinee's item responses so that it is more reasonable to assume that a specified individual either has acquired the necessary concepts and/or skills to respond correctly to (a) all items within the domain or (b) none of the items within the domain. If this kind of relation holds for the items within a domain then the only true score response patterns are 0 0 0···0 and 1 1 1···1 (i.e., the only reason why a response pattern other than all zeros or all ones occurs is due to guessing or forgetting errors on one or more of the items). Thus Model 3G (in which the only true score patterns are [0 0 0···0] and [1 1 1···1] and for each item "i" α_i and β_i are respectively guess and forget errors) and Model 3 (which is a special case of Model 3G in which $\alpha_i = \alpha$ and $\beta_i = \beta$ for all i) are appropriate models for the assessment of the nature and relations among items within domains.

In this DRT example, the item scores for 142 randomly selected students from the original sample of 284 were used to generate maximum likelihood estimates of the parameters and their standard errors under both Models 3 and 3G (presented in Table V). The data for the remaining 142 students were used as a cross-validation sample.

Note that the estimated α's , under both models, are relatively small in magnitude (except $\hat{\alpha}_1$) when compared to their standard errors and the corresponding $\hat{\beta}$-values. In fact, α_1 is the only guess error parameter that differs significantly from zero. This outcome was expected since the items were presented in free response format.

On the basis of the parameter estimates presented in Table V, expected

Table V

Maximum Likelihood Parameter Estimates and their Standard Errors for DRT Data

Model 3G			Model 3		
Parameter	Estimated value	Std. error	Parameter	Estimated value	Std. error
$\bar{\theta}$	.41	.063	$\bar{\theta}$	.40	.068
α_1	.21	.067	α	.08	.036
α_2	.07	.062			
α_3	.02	.029			
α_4	.05	.053			
β_1	.25	.059	β	.34	.041
β_2	.22	.062			
β_3	.57	.063			
β_4	.29	.065			

Table VI

Response Frequencies and Tests of Model Fit for DRT Data

Response Pattern	Observed Freq.		Expected Freq.	
	Validation sample	Cross-val. sample	Model 3G	Model 3
0 0 0 0	41	41	41.04	41.07
1 0 0 0	13	12	12.91	5.95
0 1 0 0	6	10	5.62	5.95
0 0 1 0	1	3	1.30	5.95
0 0 0 1	4	3	4.04	5.95
1 1 0 0	7	8	8.92	4.68
1 0 1 0	3	1	1.93	4.68
1 0 0 1	6	2	6.13	4.68
0 1 1 0	2	2	2.08	4.68
0 1 0 1	5	5	6.61	4.68
0 0 1 1	4	1	1.42	4.68
1 1 1 0	7	8	6.19	8.32
1 1 0 1	23	16	19.74	8.32
1 0 1 1	1	4	4.22	8.32
0 1 1 1	4	6	4.90	8.32
1 1 1 1	15	20	14.95	15.82

frequencies corresponding to each of the 16 possible response patterns were generated. These expected frequencies along with the observed frequencies for both the validation and cross-validation samples are presented in Table VI. These frequencies were used in the statistical assessment of fit provided by the models.

Table VII presents results of chi-square tests used in assessing both absolute and relative fit provided by Models 3 and 3G. Chi-square results related to model validation and cross-validation suggest reasonable absolute fit only for Model 3G.

The chi-square test related to relative fit provided by the two models resulted in significantly better fit for Model 3G. This may be interpreted as evidence supportive of the contention that: $\alpha_i \neq \alpha$ and/or $\beta_i \neq \beta$ for all i values. The large estimated value for β_3 appears to be a logically unreasonable estimate, this suggests the possible need for subdividing or otherwise restructuring this domain.

Note that it may be desirable to classify examinees obtaining each response pattern "j" in such a way that misclassification of the two true score "types" (0 0 0 0 and 1 1 1 1 which could be dubbed respectively "non-masters" and "masters") is minimized.

Given that the models are "adequate" representations of the behavior being assessed, placement may be implemented by comparing the relative magnitudes of the estimated joint proportions for each response pattern "j" with each true score type (i.e., $P(j \wedge 0000)$ and $P(j \wedge 1111)$ which are presented in Table VIII and classifying examinees obtaining response pattern "j" as:

(a) "masters" if $P(j \wedge 0000) \leq P(j \wedge 1111)$

or

(b) "non-masters" if $P(j \wedge 0000) \geq (j \wedge 1111)$.

Under this strategy for Model 3G, the response patterns designating "non-mastery" status are: 0 0 0 0, 1 0 0 0, 0 1 0 0, 0 0 1 0 and 0 0 0 1 which results in an expected proportion of misclassified examinees of .0703. For Model 3,

Table VII

Statistical Tests of Model Fit for DRT Data

Assessment	Model 3G	Model 3
Model Validation		
Chi-Square	9.459	51.758
Degrees of Freedom	6	12
P-Value	.149	.000
Model Cross-Validation[a]		
Chi-Square	12.997	34.173
Degrees of Freedom	15	15
P-Value	.603	.003
Comparison of Models		
Chi-Square	52.643	
Degrees of Freedom	6	
P-Value	.000	

[a]Model cross-validation was based on fit provided by the original expected frequencies to the observed frequencies obtained from the 142 students not used in parameter estimation.

Table VIII

Estimated Joint Proportions of Response Patterns and Mastery States for DRT Data

Response Pattern	Model 3G		Model 3	
	$\hat{P}(j \wedge 0000)$	$\hat{P}(j \wedge 1111)$	$\hat{P}(j \wedge 0000)$	$\hat{P}(j \wedge 1111)$
0 0 0 0	.2837	.0053	.2808	.0084
1 0 0 0	.0748	.0161	.0259	.0160
0 1 0 0	.0208	.0188	.0259	.0160
0 0 1 0	.0052	.0040	.0259	.0160
0 0 0 1	.0156	.0128	.0259	.0160
1 1 0 0	.0055	.0573	.0024	.0306
1 0 1 0	.0014	.0123	.0024	.0306
1 0 0 1	.0041	.0391	.0024	.0306
0 1 1 0	.0004	.0142	.0024	.0306
0 1 0 1	.0011	.0454	.0024	.0306
0 0 1 1	.0003	.0097	.0024	.0306
1 1 1 0	.0001	.0435	.0002	.0583
1 1 0 1	.0003	.1387	.0002	.0583
1 0 1 1	.0001	.0297	.0002	.0583
0 1 1 1	.0000	.0345	.0002	.0583
1 1 1 1	.0000	.1053	.0000	.1114

the same classification decisions are reached as above, however the expected proportion of misclassified examinees is .0876.

REFERENCES

Dayton, C. M. & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin & W. J. Popham (Eds.), Problems in Criterion-Referenced Measurement. Los Angeles: Center for the Study of Evaluation, U.C.L.A., 1974, 98-115.

Macready, G. B. The structure of domain hierarchies found within a domain referenced testing system. Educational and Psychological Measurement, 1975, 35, 583-598.

Macready, G. B. & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, (In press).

Macready, G. B. & Merwin, J. C. Homogeneity within item forms in domain referenced testing. Educational and Psychological Measurement, 1973, 33, 351-360.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1965.

Stouffer, S. A. & Toby, J. Role conflict and personality. The American Journal of Sociology, March, 1951, 56, 395-406.

Torgerson, W. S. Theory and Methods of Scaling. John Wiley & Sons, New York, 1958.